PROBABILISTIC ANALYSIS OF SIMPLE RANDOM GRAPH MODELS

JACOB A. SHKROB
FEBRUARY 5, 2020

Abstract. In this paper, we introduce the standard definitions for subgraphs, connected subgraphs, degree distributions, giant connected components, cliques, diameter, and typical distance of a random graph. We introduce the relevance of network theory in modern application and discuss the structures of the Erdős–Rényi random graph model and both the Non-preferential and Preferential Attachment models (NPA, PA). In this overview, we discuss bounds for the expected number of cliques of size k for $G(n,p)$ and consider the phase transition and diameter of $ER_n(p(n))$ for small p. Finally, we discuss heuristics and conjectures made about the NPA and PA models.

Contents

1. Introduction and Preliminaries

Many real-world social and hierarchical networks are based on irregular behavior and random processes, in terms of establishing connections between like individuals and objects. Network theory is a field of combinatorics that studies the behavior of such random network models, with the hope that mathematical intuitions and insights made in this paper will relate to the underlying nature of social interactions in general.

First, we shall go through notation and definitions used frequently in this essay. We denote a finite graph as $G = (V, E)$, where $V = [n] = \{1, 2, 3, ..., n\}$ are the vertices and $E = \{\{i, j\} : i, j \in V\}$ is the *edge set* of the graph G. We will use i, j and $e_{i,j}$ interchangeably in the paper. In particular, we write $G(n, m)$ to be a graph selected uniformly at random from the set of all graphs containing n vertices

and m edges. The *degree of node* i is defined as follows:

$$deg_G(u) = deg(u) = |\{v \in V : \{u, v\} \in E\}|$$

For each vertex $i \in V$, we write the degree of i as d_i, or as $d_i(t)$ if we want to emphasize the degree at time t. The *distance* between two vertices u and v is written as $dist_G(u, v)$ and denotes the length of the shortest path between them. We'll be interested in *typical distances* of a random graph model, a measure of a random graph's expected distance. We define the *diameter of* G to be the "longest distance between any two vertices in G" i.e.

$$diam(G) = \max_{u,v \in [n]} dist_G(u, v)$$

Here are some more important terms in graph and network theory used throughout the essay:

- A graph G is *complete* if every two vertices in G are adjacent.
- A *subgraph* of $G = (V, E)$ is a graph whose vertex and edge sets are subsets of V and E.
- A *clique* of a graph G is a complete subgraph of G.
- A *connected component* of a graph G is a subgraph of G in which each vertex of the component is connected to another vertex within the component.
- A *giant (connected) component* of G is a large connected component of a random graph.

We will also be using asymptotic notation to denote certain bounds of a variable as a function of the size of the graph n, such as $\Theta(n)$, $O(n)$, and $\Omega(n)$.

Throughout the paper, we will analyze the behavior of a few famous random graph models and derive certain *probabilistic* properties that they have. For example, we will consider the following kinds of questions that are important when discussing a random graph model:

- What are the expected number of edges?
- What is the degree distribution of the random graph? What is the expected degree of a randomly selected vertex?
- What is the expected number of a certain subgraph in the graph? Largest expected clique in the random graph?
- Is there a GC in the random graph?
- What is the expected diameter of the random graph?

2. Erdős–Rényi Random Graph Model

2.1. Degree distribution.

Definition 2.1. Erdős–Rényi graphs, written as $G(n, p)$, are randomly-generated graphs of n vertices such that between each pair of vertices i, j, there is a fixed probability p that an edge forms between i and j, independently from all other edges. This parameter p is called the **edge probability** (p is defined in $G(n, p)$) [2].

Consider $G(n, p)$, where $p = \mathbb{P}(i \to j)$. First, let us answer the question about the average degree of a node in $G(n, p)$. Suppose we observed a graph $G \sim G(n, p)$. Let random variables $X_i = d_i = deg(i)$ from $i = 1, 2, ..., n$ with the support $\mathcal{S}_{X_i} =$

$\{0, 1, 2, 3, ..., n-1\}$ defined by the possible degrees that node i could take on. Thus, we have that the r.v. $X_i \sim Bin(n-1, p)$ for each $i \in [n]$, where,

$$\mathbb{P}(X_j = i) = \binom{n-1}{i} p^i (1-p)^{n-1-i} \implies \mathbb{E}(X_i) = (n-1)p$$

Since we can write the total node degree as the r.v. $Y = \sum_{i=1}^{n} X_i$, we have that

$$\mathbb{E}(Y) = \mathbb{E}(\sum_{i=1}^{n} X_i) = n(n-1)p$$

Thus, the expected average degree of the nodes in $G(n, p)$ is $(n-1)p$. To find the expected number of edges in $G(n, p)$, note that each the existence of an edge accounts for a 1-degree increase in each of its vertices. Thus, we have that the expected number of edges X is exactly half of the total degree, i.e. $\mathbb{E}(X) = \frac{n(n-1)}{2} = \binom{n}{2}p = \Theta(n^2)$.

2.2. **Complete Subgraphs of $G(n, p)$.** Similarly as before, we can derive a general formula for the expected number of complete subgraphs of size l for some $l \in \mathbb{N}$. For example, consider finding the expected number of cliques of size 3. Let r.v. $T = \sum_{i,j,k} T_{ijk}$ be the total number of cliques of size 3, where

$$T_{ijk} = \begin{cases} 1 & e_{ij}, e_{jk}, e_{ki} \in V \\ 0 & \text{otherwise} \end{cases}$$

Since each edge attaches independently, $\mathbb{E}(T_{ijk}) = p^3$. Therefore, the expected number of cliques of size 3 will be

$$\mathbb{E}(T) = \mathbb{E}\left(\sum_{i,j,k \in [n]} T_{ijk}\right) = \binom{n}{3} \mathbb{E}(T_{ijk}) = \binom{n}{3} p^3$$

In general, we have that the expected number of complete subgraphs K_l in $G(n, p)$ will be $\binom{n}{l} p^{\binom{l}{2}}$.

Similarly, we can find the expected number of *non-edges* that may appear; in other words, we can find the expected number of l-empty sets of vertices in $G(n, p)$, which are subgroups of l vertices in which none are connected, using a similar approach. Instead of applying p, we use $(1-p)$ applied $\binom{l}{2}$ times. Thus, we also have that the expected number of l-empty sets of vertices in $G(n, p)$ is $\binom{n}{l}(1-p)^{\binom{l}{2}}$.

Besides the results listed above, we can find even more intriguing properties of $G(n, p)$. Instead of asking what the expected number of cliques of size l are in an ER graphs, we could consider how large the parameter l needs to be in order to observe approximately one fully connected subgraph. In order words,

Question 2.2. *For fixed n, p, find $l \in \mathbb{N}$ such that*

$$\mathbb{E}(\#\ \textit{of subgraphs}\ K_l) = \binom{n}{l} p^{\binom{l}{2}} \approx 1, \textit{ for some } l \in N$$

.

Lemma 2.3. *Proof based on results of* [1]. *For fixed $n, l \in \mathbb{Z}^+$*

$$\left(\frac{n}{l}\right)^l \leq \binom{n}{l} \leq \left(\frac{ne}{l}\right)^l$$

Proof. We can write the left-hand inequality as

$$\binom{n}{l} = \frac{n}{l} \cdot \frac{n-1}{l-1} \cdot \ldots \cdot \frac{n-(l+1)}{1} \geq \left(\frac{n}{l}\right)^l$$

.

For the right inequality, we'll use **Stirling's approximation**, which states that for asymptotically large n, $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$. Using this approximation for factorials, we can write the right-hand inequality as

$$\binom{n}{l} \leq \frac{n^l}{l!} \approx \frac{n^l}{\sqrt{2\pi l}(\frac{l}{e})^l} \leq \left(\frac{ne}{l}\right)^l$$

. □

Using these approximations, we can find asymptotic bounds for l; since $\binom{l}{2} \leq \frac{l^2}{4}$,

$$1 \approx \binom{n}{l} p^{\binom{l}{2}} \geq \binom{n}{l} p^{\frac{l^2}{4}} \geq \left(\frac{n}{l}\right)^l q^{l^2}, \text{ where } q = p^{l/2}$$

$$1 \geq \left\{ \left(\frac{n}{l}\right)^l (q^l)^l \right\}^{1/l} \implies l \geq nq^l = np^{\frac{l}{4}}$$

We can write l as $l(n)$, since l is dependent on n and substitute $p^{\frac{l}{2}}$ back into the expression. After taking the log of both sides of the inequality, ($\log(x)$ is monotonically increasing over $\mathbb{R}^+$), we get that:

$$\log(l(n)) \geq \log(n) + \frac{l(n)}{4}\log(p)$$

$$\frac{l(n)}{4} \leq \frac{\log(l(n))}{\log(p)} - \frac{\log(n)}{\log(p)}$$

Therefore, it is clear that $l(n)$ is bounded by $\log(n)$, i.e.

$$\boxed{\text{l(n) = O(log(n))}} \tag{2.4}$$

Similarly we can derive an asymptotic lower bound for $l(n)$ using Stirling's approximation from *Lemma 3.3*;

$$\binom{l}{2} \leq \left(\frac{l \cdot e}{2}\right)^2 = \frac{e^2}{4} \text{ (by Lemma 3.3)}$$

$$1 \approx \binom{n}{l} p^{\binom{l}{2}} \leq \frac{(ne)^l}{l^l} q^{l^2}, \text{ where } q = p^{\frac{e^2}{4}} \implies l \leq neq^l$$

$$\boxed{\text{l(n) = }\Omega\text{(log(n))}} \tag{2.5}$$

By equations (3.4) and (3.5), we have shown that for fixed n, p in $G(n,p)$,

$$l(n) = \Theta(\log(n))$$

where $l(n) =$ size of the largest clique in $G(n,p)$. Thus the largest "set" of individual nodes within an Erdős–Rényi random graph model, for fixed edge probability p, grows proportionately to the log of the graph's size.

2.3. **Sparse Graphs and Phase Transitions.** In the previous sections, we considered $G(n,p)$ for fixed parameter p. When varying the parameter p, it makes sense that the density of the random graph should change: for example, when $p=0$, the graph is empty and when $p=1$, the graph is complete (every edge exists with certainty). Does there exist some critical $p \in [0,1]$ such that the "structure" of $G(n,p)$ changes from sparse to dense? When a network changes its structure, it is called a *phase transition* of the network. We will call the probability necessary to observe this change the critical probability p_c. Note that when there is a phase transition, a large fraction of vertices belong to the giant component of the network. Before we calculate p_c, we need to mention sparse graphs.

Definition 2.6. A *sparse graph* of $G(n,p)$ is a regime of $G(n,p)$ where we have $p \to 0, n \to \infty$, such that $np \to \lambda \in \mathbb{R}$ [2]. In such regimes, the degree distribution of $G(n,p)$ will converge in distribution from a binomial to a Poisson random variable:

$$\mathbb{P}(d_i = k) \sim Binom(n-1,p) \xrightarrow[p\to 0]{n\to\infty} Pois(\lambda)$$

Let γ = the fraction of vertices in $G(n,p)$ belonging to the GC of $G(n,p)$. Consider the event $[v \notin GC]$ for some vertex v, and let's make the assumption that $\gamma \approx \mathbb{P}(v \notin GC)$ and that we are under a *sparse* regime. Then, using the Law of Total Probability,

$$\begin{aligned}\gamma = \mathbb{P}(v \notin GC) &= \sum_{i=0}^{n} \mathbb{P}(d_v = i)\mathbb{P}(v \notin GC | d_v = i) \\ &= \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!}\gamma^i \\ &= e^{-\lambda}e^{\lambda\gamma} = 1 - e^{\lambda(\gamma-1)}\end{aligned}$$

Thus, the number fraction of nodes belonging to GC $\rho = 1-\gamma$ can be written as the following:

$$\rho = 1 - e^{\lambda(\gamma-1)} = 1 - e^{\lambda\rho}$$

Looking closely at this equation, we can see that for $\lambda < 1$, solutions exist. For $\lambda \leq 1$, the equation has 0 solutions for all s. In particular, for $\lambda = 1$, we have no solutions, yet for $\lambda = 1+\epsilon$, solutions exist: Therefore, we can write our critical lambda $\lambda_c = 1 = np_c \implies p_c = \frac{1}{n}$.

2.4. **Diameter of $G(n,p)$.** Consider a graph $G \sim G(n,p(n))$ under the aforementioned sparse regime. We want to show that under a sparse regime, the diameter of an ER random graph approaches 0 for large enough n. In other words, we want to make the following proposition:

Proposition 2.7.

$$\mathbb{P}(dist(i,j) > 2) \to 0 \text{ as } n \to \infty$$

Furthermore,

$$\mathbb{P}\left(\bigcup_{i,j} dist(i,j) > 2\right) \to 0$$

Proof. For each $k \in [n], k \neq i,j$, let rv $X_k = 1$ if $e_{i,k}, e_{k,j} \in V$ and $X_k = 0$ otherwise, (X_k indicates when vertices i,j are connected by a path of length 2 through k). By

definition of $G(n,p)$, we know that observing 2 arbitrary connections has probability p^2, so

$$\mathbb{P}(X_k = 0) = 1 - p^2$$
$$\mathbb{P}\left(\bigcap_k X_k = 0\right) = (1-p^2)^{n-2}$$

Thus, we can now define the probability that distance between two nodes exceeds 2 and consider $G(n,p)$ as $n \to \infty$:

$$\mathbb{P}\{dist(i,j) > 2\} = \mathbb{P}(i \not\to j) \cdot \mathbb{P}\left(\bigcap_k X_k = 0\right) = (1-p)(1-p^2)^{n-2} \approx \alpha^n, \alpha < 1$$

$$\mathbb{P}\left\{\bigcup_{i,j} dist(i,j) > 2\right\} \leq \sum_{i,j \in [n]} \mathbb{P}\{d(i,j) > 2\} = \binom{n}{2}\alpha^n \approx n^2\alpha^n \to 0, \alpha < 1$$

Which holds for fixed p. □

If $p(n) \to 0$ (sparse random graph), we can write $\mathbb{P}(dist(i,j) > 2)$ as

$$\mathbb{P}(dist(i,j) > 2) = (1-p)^2(1-p^2)^{n-2} = \left\{\frac{(1-p)}{(1-p^2)^2}\right\}(1-p^2)^n \overset{p \ll 1}{\approx} (1-p^2)^n$$

We'll make the assumption that if $p(n) \to 0$, $p^2 \simeq \frac{r}{n}$ for some $r \ll n$. Then, we know that

$$(1-p^2)^n \to (1-\frac{r}{n})^n \underset{n\to\infty}{\to} e^{-r}$$

We now ask the following question: for what regime of the ER random graph does the probability approach 1? In other words, we ask for which values of $p(n)$ allow our graph G to have, with probability ≈ 1, the $diam(G) > 2$? If the event $\{diam(G) > 2\}$ occurs for graph $G \sim G(n,p)$, surely we know that the event $\left\{\bigcup_{i,j} dist(i,j) > 2\right\}$ must also occur. Thus, we have the following:

$$1 \approx \mathbb{P}\left\{\bigcup_{i,j} dist(i,j) > 2\right\} \leq \sum_{i,j \in [n]} \mathbb{P}\{d(i,j) > 2\} \leq n^2(1-p(n)^2)^n \underset{n\to\infty\ p\to 0}{\longrightarrow} n^2 e^{-p^2 n}$$

$$\lim_{n\to\infty} n^2 e^{-p^2 n} = 1$$

$$\implies \lim_{n\to\infty} \log(n^2 e^{-p^2 n}) = \lim_{n\to\infty} (2\log(n) - p^2 n) = \log(1) = 0$$

$$\implies \lim_{n\to\infty} \left(\frac{2\log(n)}{p^2 n}\right) = 1$$

In summary, we have that for large n, $p^2 n \approx 2log(n)$, meaning asymptotically,

$$p^2 = \Theta\left(\frac{\log(n)}{n}\right)$$

In general, we have found the following rule: if $p^2 = \Theta\left(\frac{\log(n)}{n}\right)$, then indeed, the diameter of $G \sim G(n,p)$ is almost surely *greater than* 2. Otherwise, the diameter of G will be ≤ 2 or potentially infinite if $p(n)$ quickly goes to 0. We have bounded the thresh-hold for our edge parameter $p(n)$.

3. Introduction to the Non-Preferential Attachment Model

The Non-Preferential Attachment Model (NPA) is a more interesting abstraction of a randomly-generated graph; in order to model networks growing and evolving with time, such as social networks and citation networks, we have to create a random graph model that *grows* with time. These kinds of random graph models are called **stochastic growth models**.

Definition 3.1. The **NPA model** starts with a fully-connected graph of K vertices, for some $k \in \mathbb{N}$. At time $t = 0$, we create the fully-connected subgraph of size k. At time $t = 1, 2, 3, ...$, a new vertex is introduced to the graph that can make at most k edges with the present vertices; in other words, the i-th introduced vertex has an initial degree of k at time $(d_i(i) = k)$. Lastly, for each vertex introduced to the graph, it will form k edges uniformly at random with the remaining nodes of a graph i.e.

$$\mathbb{P}(i \to j) = \frac{1}{K + t - 1} \quad (t = 1, 2, 3, 4, ...)$$

Consider the simplest example where $K = 1$. Then, it is clear that the expected degree of the first node introduced $d_i(t)$ will be

$$\mathbb{E}(d_1(t)) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + ... + \frac{1}{t} = \sum_{i=1}^{t} \frac{1}{i} \approx \log(t)$$

Similarly, if we start at time $t = i$, then the expected degree of the i-th node introduced will have a similar value to the first vertex, only shifted slightly:

$$\mathbb{E}(d_i(t)) = 1 + \frac{1}{i+1} + \frac{1}{i+2} + ... + \frac{1}{t} \to \log(t) - \left(\frac{1}{2} + \frac{1}{3} + ... + \frac{1}{i}\right) \to log\left(\frac{t}{i}\right)$$

If we instead consider NPA model for some general K, k, then in a similar way, for the i-th node introduced to the network,

$$\mathbb{E}(d_i(t)) = k + k \cdot \frac{1}{i+K} + k \cdot \frac{1}{K+i+1} + ... + k \cdot \frac{1}{K+t-1} \overset{t \gg K}{\longrightarrow} k\left(1 + log\left(\frac{t}{i}\right)\right)$$

Conjecture 3.2. *Although not proven, we can try to give an argument as to what the typical distance of the NPA model could be. For our simplistic model with $K = 1$, suppose we take our assumption from before that $d_i(t) = O(\log\left(\frac{t}{i}\right))$. If we take a fixed vertex a, and make the assumption that there is a connection from vertex a to vertex $\approx \frac{a}{2}$, then $\mathbb{E}(dist(0, a)) \approx \log(a)$. This means that in the worst case scenario, such a model would have an expected diameter of about $2\log(n)$, since we have 2 different paths of length approximately $\log(n)$ between the furthest vertices.* [3]

This concludes the section on the Non-Preferential Attachment model. In the future, we hope to do further research into the Preferential Attachment model and understand the nature of its diameter and typical distance. A comparison between the simpler non-preferential model and the more complex preferential attachment model would make for a very interesting research topic. Finally, research on the configuration model and its possible centrality measures would be a worthwhile topic to investigate.

Acknowledgments

It was a pleasure to a part of The Department of Mathematics and Statistic's second iteration of the Undergraduate Directed Reading Program. Thank you to my mentor, Jordan Barrett, for spending time with me and teaching me an exciting application of graph theory and probability. Thanks so much for spending so much time with me, it was incredibly fun learning with you.

4. Bibliography

References

[1] Das, Shagnik. A brief note on estimates of binomial coefficients. (2016)
[2] R. van der Hofstad. Random Graphs and Complex Networks. Volume 1. Cambridge Series in Statistical and Probabilistic Mathematics (2017)
[3] Dommers, Sander, Remco van der Hofstad, and Gerard Hooghiemstra. "Diameters in Preferential Attachment Models." Journal of Statistical Physics 139.1 (2010): 72–107. Crossref. Web.
[4] P. Erdos and A. Renyi. On random graphs I. Publ. Math. Debrecen, 1959.